Stat 104 Spring 2016
Regression Project
Due 4pm May 3, 2016

Collaboration.

You must work by yourself on this project. No collaboration with anyone else (in this class or not) is permitted. You may consult with the teaching staff as the need arises.

Comment on the role of your TF

Please only consult with your TF when you run into a problem, not when you are just looking to have someone check your work. Otherwise, contacts with your TF may become excessive and this project tends to become a "joint project with your TF" rather than "your own project."

Overview

An individual "regression report" is required of all students. Your regression report is **due before 4:00 pm, May 3, 2016**. Submit the report electronically on Canvas as you would a homework assignment. Each student will analyze the same data set, described below. The report is expected to be 5-10 pages in length.

Project Background

Imagine you are the owner(s) of a company, CrimsonCream, Inc. that operates stores in New York, Chicago, and Los Angeles. You have been operating for quite a long time, and believe you understand your market well. Not long ago, you observed your market share beginning to slip a bit and decided to embark on a promotion campaign in all three cities to help to increase your sales. At this point, after the marketing campaign is over, you would like to investigate whether it worked. Your in house economist (Greg Mankiw) has collected the data listed below for you, for the 365 days of the year 2016. Yes, as many have thought, Greg owns a time machine.

Your assignment is to do the following:

a) Summarize the data in this data set in a way that will help the reader of your report see a basic picture of your ice cream business. Use standard statistical devices, including graphs. Make this a concise readable, summary of the data that will help your reader understand the analysis to follow.

b) Determine, using regression methods, what effects that are contained in your data help to explain your ice cream sales. That is, specify and estimate an equation that adequately describes your sales. Present the relevant statistical results in neat, understandable tables that will enable your reader to learn about your market from your model.

c) Answer the question "Did the promotional campaign work?" statistically. Justify your answer with the results of your estimated model.

Project Data

For each city, you have collected data on each of the 365 days of 2016. The following variables are in your database:

Three cities' populations: New York, 8 million, Chicago, 4 million, Los Angeles, 3 million

- SALESPERCAP = Total sales per capita of cups of ice cream. This is your response variable to model using multiple regression
- PRICE = The average price that stores are charging for your ice cream. The stores differ quite a bit in the price they charge, so this variable changes quite a bit
- COMP = Your main competitor's average price. Since your competitor claims they are a premium" brand, while you are for "every person," COMP is generally higher than PRICE
- TEMP_NY, TEMP_CHI, TEMP_LA are the average temperatures in the cities each day in the year.
- ECONOMY = an indicator of overall national economic sentiment. 2016 wasn't a very good year, and it got worse as the year went on, so the index declines over the year.
- NEWCHAIN = a variable that indicates that a very large new restaurant chain that specializes in lunch, and serves your product, opened in Chicago in the middle of the year. This is a binary variable that equals one for the days when this chain was open.
- WEEKEND = a binary variable that is one on Saturday and Sunday. People like to go out for ice cream on weekends.
- LA, CHI, NY are three binary variables that indicate which observations are for which city
- DAY_WEEK = a weekday variable, taking values 1,2,3,4,5,6,7,1,2,3,4,5,6,7,…
- DAY_YEAR = a year day variable that takes values from 1 to 365, for each city.
- PROMO = a binary variable that equals 1 from July 1 until the end of the year. This is when you ran your promotion.

The data may be loaded into Stata using the command

```
use http://people.fas.harvard.edu/~mparzen/stat104/icecream2016V1
```

You will need to think about all of the technical problems that can arise in regression. For this data set multicollinearity, heteroscedasticity, and outlying observations may be problems. You will probably wish to carry out some hypothesis tests as part of your work; these will be of dubious value if the assumptions of the normal multiple linear regression model are badly violated.

The report itself must be no more than 10 pages, including graphs and any Stata output you wish to include. We want your model building process explained in an easy to follow format (this includes why variables were excluded, transformed or modified, and any diagnostics that were performed).

Evaluation

The grade on the assignment will depend on three things.

1. *Clari*ty (50%). The report must clearly indicate how you went about your work, including what models were considered.

2. *Substance* (50%). The project should use the tools developed in our class in an appropriate and correct manner. The report should anticipate questions that a technically critical reader might ask. For example, if the model can predict negative saless for certain reasonable values of the regressors, and there is no discussion of this fact, there are problems. Similarly, if heteroscedasticity is likely to be a problem with a particular function form, then the report must indicate how this was handled.

Notes:
Note that we enjoy graphs like regression diagnostic plots and scatter plots of response versus explanatory variables (at least one).

Clearly walk the reader through the analysis performed, but in a readable way without using a lot of computer output or jargon.